Motivation and Problem Definition:

Social Network websites becomes the major environment for the users to communicate and express their opinions. But, there are some users use the social network websites in negative way like cyberbullying.

The united states government defines the Cyberbullying as a bullying that takes place over digital devices like cell phones, computers, and tablets. And, it can occur through SMS, Text, and apps, or online in social media, forums, or gaming where people can view, participate in, or share content. Cyberbullying includes sending, posting, or sharing negative, harmful, false, or mean content about someone else. It can include sharing personal or private information about someone else causing embarrassment or humiliation [1]. Moreover, multiple studies mention that cyberbullying can affect any anyone especially young people and they are at high risk. For example, National Center for Education Statistics and Bureau of Justice Statistics reported that 9% of students in grades 6–12 experienced cyberbullying [2]. Moreover, Kosciw, J. G. et al. say that 55.2% of LGBTQ students experienced cyberbullying [3].

Cyberbullying has negative consequences on victim. For example, Sourander et al. say that cyberbullying leads to serious pathological experience such as depression, self-harm and suicide attempt [4]. In addition, S. Hinduja and J. W. Patchin say that the effects of cyberbullying can start from temporary anxiety to suicide [5]. Kowalski and limber mention that 90% of the young people victims don't tell their parent about their cyberbullying experience [6].

Manual detection of cyberbullying is very difficult because the huge volume of data on social network websites. Therefore, accurate and automated detection of cyberbullying is more effective. Several studies focused mainly on the content-based feature such as profanity, pronouns, bags of word, term frequency inverse document (TFIDF) and cyberbullying words. For example, Foong and Oussalah present an automated cyberbullying system detection. The system is based on natural language processing, text mining and machine learning to detect cyberbullying. The authors employed different textual features for the classifier such TF-Idf, linguistic Inquiry, word count features and Dependency features. In addition, the authors conduct their experiment on the collected dataset from ASKfm website [7]. Little studies are focused on user-based feature. Therefore, in this project, we will focus on the user-based feature and we will try to extract new user-based feature to enhance the accuracy of detection cyberbullying.

Related Works:

Silva et al.[8] present first version of bully blocker application which is designed for the parents to help them monitoring Facebook interactions of their adolescent can used to detect and alerting the parents when cyberbullying occurs. Zhang et al.[9] propose a novel pronunciation based on convolution neural network to face the noise and error in social media posts and messages make detecting cyberbullying very challenging. In addition, the authors used phoneme codes of the text as features for CNN. This procedure corrects spelling errors that did not alter the pronunciation.

Romsaiyed et al.[10] present an automatic cyberbullying detection system to detect, identify, and classify cyberbullying activities from large volume of streaming text from OSN services. Text are fed into cluster and discriminate analysis stage which can identify abusive text. Then the abusive texts are clustered by using K-Mean. Naïve Bayes is used as classification method to build classifer from our training datasets and build predictive model. Dinakar et al.[11] focus on textual cyberbullying detection. In addition, the authors used different features such as tf-idf, ortony lexicon for negative affect, list of profane words, POS bigrams: jj_DT,PRP_VBP, VB_PRP and top specific unigram and bigrams. Moreover, the authors conduct their experiment on collected data from YouTube. Hee et al.[12] present automatic cyberbullying detection in social media text by modeling posts written by bullies, victims and bystanders of online bullying. Moreover, the authors describe the collection and fine-grained annotation of a training corpus for English and Dutch and perform a series of binary classification experiments. In addition, the authors extracted different features word n-gram bag-of-words, character n-grams bag-of-words, term lists, subjectivity lexicon features and topic model features. Rafiq et al.[13] design a novel approaches to detect instances of cyberbullying over Vine media sessions which is a mobile based video sharing online social network. Moreover, the authors collect a set Vine video session and us CrowdFlower to label media session for cyberbullying and cyberaggression. Rafiq et al.[14] propose a multi-stage cyberbullying detection solution that drastically reduces the classification time and the time to raise cyberbullying alerts. Moreover, the authors proposed solution which is scalable, does not sacrifices accuracy for scalability. In addition, the solution is comprised of three novel components, an initial predictor, a multilevel priority scheduler and incremental classification mechanism. Chelmis et al.[15] perform a detailed analysis of a large-scale real world dataset to identify online social network topology structure features that are the most prominent in enhancing the accuracy of state of the art classification methods for cyberbullying detection. Moreover, the authors derived small subset of features that are fast to compute while differentiating between normal users cyberbullies and victims. Rafiq et al.[16] propose a multi-stage cyberbullying detection solution that drastically reduces the classification time and the time to raise cyberbullying alerts. Moreover, the authors proposed solution which is scalable, does not sacrifices accuracy for scalability. In addition, the solution is comprised of two novel components, dynamic priority and incremental classification mechanism. Li et al.[17] develop methods for detecting cyberbullying based on sharing images on Instagram. Moreover, the authors extracted images specific and text features from comments and from image captions. In addition, several novel features including topics determined from image captions and outputs of pretrained conventional neural network applied to image pixels. Hosseinmardi et al.[18] investigate the prediction of cyberbullying incidents in Instagram. Moreover, the authors build a predictor that can be anticipate the occurrence of cyberbullying incidents before they happen.

Methodology:

Dataset Description

In this project, we will use the dataset on cyberbullying which is collected by impermuim on Kaggle (https://www.kaggle.com/c/detecting-insults-in-social-commentary/data). The dataset is included six files. we will use the dataset in train.csv file. the total number of samples in the dataset is 3947. We will divide the dataset into 3000 samples train dataset and 947 samples test dataset.

Dataset Preprocessing

The dataset preprocessing stage includes noise removal (removing HTML, XML and metadata), tokenization means that breaking the stream of text into small units called token. Normalization which includes removing stop words, converting the uppercase into lower case, non-ASCII characters and stemming.

Feature extraction

We will choose the textual features that will be applied for classifiers. Two types of feature extraction will be applied count vector feature and TF-IDF feature.

Result and Evaluation:

In this project, we used different classification models to detect cyberbullying phenome. We evaluate our models by using accuracy. The accuracy is ranged between 76% -80%. The result shows that SVM classifier has the highest accuracy among other classifiers. The results of classifiers are displayed in the Table 1.

Classifiers	Accuracy
Perceptron	**0.72668**
SVM	**0.80147**
Decision Tree	**0.7756232**
KNN	**0.7654**
Logistic Regression	**0.765466**

Table 1: Accuracy Result by Applying different Classifiers on test data based on TF-IDF Features

References:

[1] “What is Cyberbullying” Accessed March 15,2018. [online]. Available: https://www.stopbullying.gov/cyberbullying/what-is-it/index.html

[2] National Center for Education Statistics and Bureau of Justice *Statistics*, *2011*.

[3] Kosciw, J. G., Greytak, E. A., Bartkiewicz, M. J., Boesen, M. J., & Palmer, N. A, "The 2011 National School Climate Survey: The experiences of lesbian, gay, bisexual and transgender youth in our nation's schools*". New York: GLSEN.*

[4] Sourander, A., Brunstein-Klomek, A., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen, M., Ristkari, T., Hans Helenius, H. 2010. "Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study." *Arch Gen Psychiatry*, 67: 720-728.

[5] S. Hinduja and J. W. Patchin. "Bullying, cyberbullying, and suicide". *Archives of suicide research*, 14(3):206–221, 2010.

[6] Robin M Kowalski and Susan P Limber." Electronic bullying among middle school students". *Journal of adolescent health*, 2007,41(6):S22–S30.

[7] Y. J. Foong and M. Oussalah, "Cyberbullying System Detection and Analysis," *2017 European Intelligence and Security Informatics Conference (EISIC)*, Athens, 2017, pp. 40-46.
doi: 10.1109/EISIC.2017.43

[8] Y. N. Silva, C. Rich, J. Chon and L. M. Tsosie, "BullyBlocker: An app to identify cyberbullying in facebook," *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, San Francisco, CA, 2016, pp. 1401-1405. doi: 10.1109/ASONAM.2016.7752430

[9] X. Zhang *et al*., "Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network," *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Anaheim, CA, 2016, pp. 740-745. doi: 10.1109/ICMLA.2016.0132

[10] W. Romsaiyud, K. na Nakornphanom, P. Prasertsilp, P. Nurarak and P. Konglerd, "Automated cyberbullying detection using clustering appearance patterns," *2017 9th International Conference on Knowledge and Smart Technology (KST)*, Chonburi, 2017, pp. 242-247. doi: 10.1109/KST.2017.7886127

[11] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," in Proc. IEEE International Fifth International AAAI Conference on Weblogs and Social Media (SWM'11), Barcelona, Spain, 2011.

[12] Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W. and Hoste, V. (2018). Automatic Detection of Cyberbullying in Social Media Text. [online] Arxiv.org. Available at: https://arxiv.org/abs/1801.05617 [Accessed 11 May 2018].

[13] R. Ibn Rafiq, H. Hosseinmardi, R. Han, Qin Lv, S. Mishra and S. A. Mattson, "Careful what you share in six seconds: Detecting cyberbullying instances in Vine," *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Paris, 2015, pp. 617-622. doi: 10.1145/2808797.2809381

[14]. R. Ibn Rafiq, H. Hosseinmardi, R. Han, Qin Lv, S. Mishra, "Investigating Factors Influencing the Latency of Cyberbullying Detection".

[15] C. Chelmis, D. S. Zois and M. Yao, "Mining Patterns of Cyberbullying on Twitter," *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, New Orleans, LA, 2017, pp. 126-133. doi: 10.1109/ICDMW.2017.22

[17] Zhong, H., Li, H., Squicciarini, A., Rajtmajer, S., Griffin, C., Miller, D., & Caragea, C. “Content-Driven Detection of Cyberbullying on the Instagram Social Network". in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2016.

[18] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv and S. Mishra, "Prediction of cyberbullying incidents in a media-based social network," *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, San Francisco, CA, 2016, pp. 186-192. doi: 10.1109/ASONAM.2016.7752233